Prediction of Appointment No Show

Abeer Almdani | Rana Alzahrani

Abstract

Patients make appointments at clinics or hospitals to be checked by a doctor. Some of the patients do not show up for their appointments. This results in loss of valuable resources in terms of physician time and staffing allocation which could have been used more productively.

Design

In this project, We will predict if the person shows up or not using a few classification algorithms.The aim is to help clinics in Brazil know if the clinic or the patient is responsible for this problem and understand the causes that led to it.

Data

data provided by Kaggle.

- **Data Size before Feature Engineering and UnderSampling:**
    - Number of rows: 110528
    - Number of columns: 14
- **Data Size after Feature Engineering:**
    - Number of rows:110527
    - Number of columns: 98
- **Data Size after UnderSampling:**
    - Number of rows:83803
    - Number of columns: 98
- **Splitied data: Training set:**
    - Before UnderSampling: Counter({1: 83803, 0: 21197})
    - After UnderSampling: Counter({1: 83803, 0: 83803})

- **Columns:**
  - PatientId:Identification of a patient.
  - AppointmentID:Identification of each appointment.
  - Gender:Male or Female.
  - ScheduledDay:The day the patient booked the appointment.
  - AppointmentDay:The day of the appointment.
  - Age:How old is the patient.
  - Neighbourhood:Where the appointment takes place.
  - Hipertension:True or False (High blood pressure).
  - Diabetes:True or False.
  - Alcoholism:True or False.
  - Handicap:0-4 (the handicap refers to the number of disabilities a person has. For example, if the person is blind and can't walk the total is 2)
  - SMS_received:1 or more messages sent to the patient.
  - No-show:True or False.
- **Feature Engineering Columns**
  - Moring : The ScheduledDay in the morning or not .
  - NeighbourhoodCount : count .
  - days : difference between ScheduledDay and AppointmentDay
  - Neighbourhood :dummies

Algorithms and Models

- Logistic Regression
- Polynomial degree = 2
- Knn
- Decision Tree
- Random Forest
- Extra Trees
- Grid Search Cross Validation
- Accuracy ,F1 score, Recall
- ROC score

Tools

- Sklearn
- Numpy
- Pandas
- Matplotlib
- Seaborn
- Python
- Anaconda
- Jupyter notebook
- Canva
- Github
- Zoom

Modules	AccuracyTr	Validation Score	Grid search
BaseLine	79.80	79.869	C=1
Logistic regression	79.52%	79.65%	C=1
Knn	99.36%	80.25%	n_estimators= 161
Decision Tree	86.78%	77.75%	max_depth=21
Random Forest	99.29%	79.36%	n_neighbors=55
Extra Trees	99.29%	78.04%	-
XGBClassifier	87.33%	80.25%	-
Polynomial 2	99.20%	99.09%	-

55-nearest neighbors algorithm is the best algorithm

Score of train set: 0.99358
Score of validation set: 0.8025
Score of test set: 0.79861

Standard classification metric

Accuracy: 0.79861
Precision: 0.33333
Recall: 0.00045
F1_Score: 0.0009